JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika

[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

KLASIFIKASI SENTIMEN X-TWITTER PERIHAL PEMINDAHAN IBU KOTA INDONESIA MENGGUNAKAN EKSTRAKSI FITUR TF-IDF DAN METODE SUPPORT VECTOR MACHINE (SVM)

Tri Wahyudi [a,1], Rudiman [b,2*],, Naufal Azmi Verdikha [c,3]

[a,b,c] Universitas Muhammadiyah Kalimantan Timur, Jl. Ir. H. Juanda No.15, Sidodadi, Kec. Samarinda Ulu, Kota Samarinda, Kalimantan Timur 75124
[1] 2011102441035@umkt.ac.id; [2] rudiman@umkt.ac.id; [3] nav651@umkt.ac.id
* corresponding author

ARTICLE INFO

Keywords
Klasifikasi Sentimen, Pemindahan Ibu Kota, X-Twitter, TF-IDF, Support Vector Machine (SVM)

ABSTRACT

The classification model has reached the realm of sentiment classification to analyze user sentiment in providing comments. this research aims to classify sentiment regarding the topic of moving the capital city of Indonesia using the Support Vector Machine (SVM) method with TF-IDF weighting. SVM has its own advantages, namely to overcome complex problems in SVM classification using the kernel function. the kernel functions to transform input data into a high dimensional feature space, allowing linear separation of data more easily. there are 3 sentiment categories in this study, namely Negative, Neutral and Positive sentiment. to determine these 3 categories, researchers used expert labelling services. the purpose of this study using the SVM method and TF-IDF feature extraction is to find out and analyze the accuracy results obtained in processing sentiment data regarding the transfer of the capital city of Indonesia. The accuracy results obtained are 73%, this shows that the SVM method with TF-IDF weighting is able to classify sentiment data with fairly good results.

1. Pendahuluan

Ibu kota negara, menurut definisi KBBI, merupakan pusat konsentrasi komponen pemerintahan, meliputi eksekutif, legislatif, dan yudikatif. Istilah "capitol" untuk ibu kota berakar dari kata Latin "caput" yang bermakna kepala, merujuk pada pusat pemerintahan. Jakarta secara resmi ditetapkan sebagai ibu kota negara Indonesia melalui dua regulasi penting: Undang-Undang Republik Indonesia Nomor 10 Tahun 1964 tentang Pernyataan Daerah Istimewa, dan Undang-undang Nomor 29 Tahun 2007 yang mengatur tentang pemerintahan provinsi daerah khusus Jakarta sebagai ibu kota Negara Kesatuan Republik Indonesia. Hingga saat ini, ibu kota Indonesia tetap menjadi bagian dari wilayah metropolitan Jabodetabek [1].

Pemilihan Provinsi Kalimantan Timur sebagai lokasi Ibu Kota Negara (IKN) baru merupakan

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

langkah strategis untuk mendorong pemerataan pembangunan, terutama di kawasan Indonesia timur.

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

Kalimantan dipandang sebagai pilihan ideal karena aksesibilitasnya yang baik, populasi yang besar dan terbuka, keberagaman budaya, serta risiko konflik yang rendah, menjadikannya barometer yang tepat untuk pengembangan IKN baru di wilayah timur. Keputusan ini didukung oleh berbagai faktor, termasuk ketersediaan Tri Matra (darat, laut, udara), infrastruktur yang memadai, serta sumber daya air yang melimpah dari tiga waduk yang ada, dua waduk yang direncanakan, empat sungai utama yang strategis, dan empat daerah aliran sungai penting [2]. Pemindahan IKN hanya menyumbangkan 0,02% pertumbuhan ekonomi nasional, menurut kajian Indef awal 2020. Dengan kenaikan pertumbuhan sekitar 3,14%, hanya provinsi tersebut yang mengalami dampak yang signifikan [3].

Twitter telah menjadi platform signifikan dalam pencarian informasi aktual, menawarkan beragam fitur yang memenuhi kebutuhan penggunanya. Di era kemajuan teknologi yang pesat ini, media sosial, khususnya Twitter, telah menjadi bagian integral dari kehidupan sehari-hari masyarakat. Platform ini tidak hanya populer, tetapi juga berfungsi sebagai sarana penting bagi publik untuk berbagi informasi dan mengekspresikan pendapat melalui tweet. Keunggulan Twitter terletak pada kemampuannya menyediakan akses cepat ke opini pengguna, menjadikannya sumber informasi yang kaya dan beragam, mencerminkan pandangan dan perasaan masyarakat umum tentang berbagai topik terkini [4]. Data dari Statistic menunjukkan bahwa pada Juli 2021, Indonesia menduduki posisi keenam global dalam jumlah pengguna Twitter, dengan total 15,7 juta pengguna (Dwi Pangestu et al., 2022). Dalam konteks ini, analisis sentimen menjadi alat yang sangat penting untuk menginterpretasikan dan memprediksi opini publik yang tercermin dalam tweet. Teknik klasifikasi sentimen tidak hanya membantu dalam memecahkan tantangan prediksi sentimen dari data Twitter, tetapi juga mempermudah proses kategorisasi dan analisis tweet pengguna, memberikan wawasan yang berharga tentang opini dan tren di kalangan pengguna Twitter Indonesia [5].

Klasifikasi opini melalui pendekatan text mining membutuhkan metode yang akurat dan efektif, seperti Support Vector Machine (SVM). SVM unggul karena kemampuannya menggunakan fungsi kernel untuk mengimplementasikan hyperplane input nonlinier berdimensi tinggi. Metode ini menawarkan solusi bernama kernel untuk mengatasi permasalahan kompleks dalam klasifikasi. Kernel berfungsi mentransformasikan data input ke dalam ruang fitur berdimensi tinggi, memungkinkan pemisahan data secara linier dengan lebih mudah. Keunggulan ini memungkinkan SVM untuk menangani data dengan struktur yang tidak beraturan, meningkatkan efektivitas klasifikasi dalam konteks analisis sentimen dan opini [6].

Pada beberapa penelitian sebelumnya banyak peneliti yang menggunakan metode SVM untuk melakukan klasifikasi sentimen,diantaranya peneliti pertama seperti yang dilakukan oleh [7]. Analisis sentimen pada maskapai penerbangan di platform twitter menggunakan algoritma Support Vector Machine. Akurasi rata-rata yang diperoleh adalah sebesar 80,41%. Berdasarkan dari penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma (SVM) terbukti mampu diterapkan dengan baik untuk melakukan analisis sentimen dengan didukung oleh beberapa metode preprocessing, pembobotan term menggunakan TF-IDF, dan parameter tuning algoritma. Kemudian pada penelitian [8] mempelajari analisis sentimen di Twitter mengenai calon presiden tahun 2019 dengan menggunakan teknik Support Vector Machine (SVM). Tujuannya untuk mengetahui besarnya keyakinan masyarakat terhadap pemilu presiden Indonesia tahun 2019.

Untuk mengatasi permasalahan ketidakseimbangan data maka dalam penelitian ini menggunakan Teknik oversampling. Oversampling sendiri mencakup metode seperti Random Over Sampling(ROS),

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

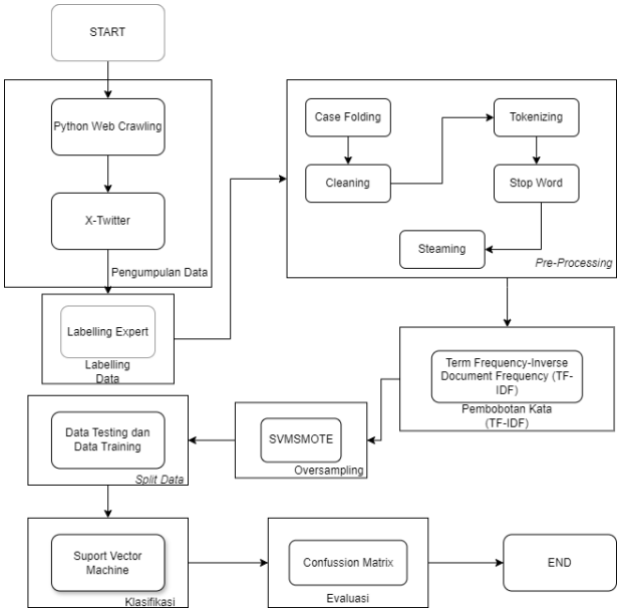
JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

Synthetic Minority Oversampling Technique(SMOTE), Borderline SMOTE, k-Means SMOTE, Support Vector Machine SMOTE(SVM-SMOTE), dan Adaptive Synthetic. Pada penelitian ini, digunakan model SVM-SMOTE karena efektif dalam menangani ketidakseimbangan kelas dan mengurangi overfitting, serta menghasilkan akurasi yang baik.

Tujuan penelitian ini adalah untuk menganalisis sentiment Masyarakat Indonesia terhadap pemindahan Ibu Kota Negara ke Kalimantan Timur menggunakan data dari platform Twitter. Tahapan penelitian meliputi pengumpulan data tweet, pre-processing text, pembobotan term menggunakan TF-IDF, dan klasifikasi sentiment menggunakan metode Support Vector Machine (SVM). Dengan menggunakan teknink ini, diharapkan dapat diperoleh wawasan yang akurat tentang opini publik mengenai pemindahan ibu kota. Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan metode klasifikasi sentiment yang efektif untuk isu-isu kebijakan publik di Indonesia, serta meningkatkan pemahaman tentang peran media sosial dalam membentuk opini Masyarakat.

2. Metodologi Penelitian

Penelitian ini melibatkan serangkaian tahapan metodologis yang sistematis. Dimulai dengan proses pengumpulan data, dilanjutkan dengan tahap preprocessing untuk mempersiapkan data mentah. Selanjutnya, metode TF-IDF diimplementasikan guna menghitung bobot kata-kata kunci. Setelah itu, data dibagi (split) menjadi set pelatihan dan pengujian. Akhirnya, proses klasifikasi dilakukan menggunakan algoritma Support Vector Machine (SVM) untuk menganalisis dan mengkategorikan data yang telah disiapkan. [9]



Gambar 10 Alur Penelitian

Penelitian ini melibatkan serangkaian tahapan metodologis yang dimulai dengan pengumpulan data atau crawling dari platform media sosial Twitter (X). Selanjutnya, data melalui proses labelling dan

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

preprocessing yang mencakup Case folding, cleaning data, Tokenizing, Stopword removal, dan stemming. Setelah itu, data latih diberi label dan dilakukan pembobotan kata menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency). Sebelum ke tahap split data metode yang harus dilakukan adalah metode oversampling menggunakan svmsmote. Tahap akhir melibatkan pelabelan dan klasifikasi data yang telah diproses menggunakan algoritma Support Vector Machine. Rangkaian langkah ini dirancang untuk mempersiapkan dan menganalisis data secara komprehensif guna mencapai tujuan penelitian. [10].

2.1. Pengumpulan Data

Proses pengumpulan data meliputi pengumpulan data yang dikumpulkan selama proses tersebut. Dalam pengumpulan data Twitter (X) [9] digunakan tools dari Google Collab serta kumpulan data diambil melalui twitter API untuk mendapatkan Twitter Token yang mana dapat digunakan untuk menarik data tweet, kumpulan data dalam format file CSV.

2.2. Labelling Data

Dalam penelitian ini, penulis memanfaatkan layanan pelabelan data melalui platform marketplace freelance https://projects.co.id/. Situs ini berfungsi sebagai penghubung antara penyedia jasa dan klien yang membutuhkan jasa atau produk digital. Untuk tugas pelabelan, penulis mempekerjakan seorang profesional freelance yang sudah memiliki ketrampilan dalam melakukan labelling data sentiment.

2.3. Pre-Processing

Preprocessing merupakan langkah krusial dalam analisis sentimen, khususnya untuk data tidak terstruktur seperti tweet. Proses ini bertujuan mengolah data menjadi format yang lebih mendasar, umum, dan standar, sehingga dapat dianalisis dengan efektif. Tahap ini menjadi sangat penting, terutama untuk tweet dalam bahasa daerah yang perlu diterjemahkan ke bahasa Indonesia. Dalam penelitian ini, preprocessing dilakukan pada data training melalui serangkaian tahapan yang terstruktur dan berurutan, memastikan data siap untuk analisis lebih lanjut.

a. Case Folding : langkah untuk mengubah teks dokumen menjadi bentuk standar, khususnya huruf kecil. Contoh dari Case Folding, yaitu “TAS” menjadi “tas”, “GELAS” menjadi “gelas”, “DASAR” menjadi “dasar” dan lain-lain.
b. Cleaning : proses menghilangkan elemen tertentu yang terdapat pada tweet, khususnya Uniform Resource Locator (URL), username, RT (Retweet), karakter HTML, dan hastag.
c. Tokenizing : pada tahap ini, sebuah tweet atau komentar dari pengguna yang sudah melalui tahap cleaning dan case folding dipisahkan dari kalimatnya menjadi sebuah kata, juga dapat menafsirkan dan mengelompokan token yang terisolasi untuk membuat token dengan tingkat yang lebih tinggi.
d. Stopword Removal : kata-kata yang tidak memiliki makna arti yang jelas seperti imbuhan akan dihilangkan dari data yang digunakan pada penelitian. Contohnya seperti ingin, di, adalah, namun, mereka dan sebagainya. Dibawah merupakan contoh dari code stopword removal yang menggunakan kamus Sastrawi.
e. Stemming : proses penguraian kata menjadi kata dasar tanpa imbuhan di awal atau akhir kata, yaitu dengan menghilangkan awalan dan akhiran kata.

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

2.4. Pembobotan Kata (TF-IDF)

Pembobotan TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode untuk mengkonversi data teks ke dalam format numerik dan memberikan bobot pada setiap kata atau fitur. Teknik statistik ini digunakan untuk menilai signifikansi kata dalam suatu dokumen. TF mengukur frekuensi kemunculan kata dalam setiap dokumen, menunjukkan relevansi kata tersebut, sementara DF menghitung jumlah dokumen yang memuat kata tersebut, mengindikasikan seberapa umum kata itu digunakan. Tujuan utama TF-IDF adalah mengidentifikasi kata-kata kunci dalam dokumen atau kumpulan dokumen. Frequency Term (TF) merujuk pada frekuensi kemunculan kata dalam dokumen, yang dapat dihitung menggunakan rumus tertentu.

$$tf_{t.d} = \frac{Jumlah\ kemunculan\ kata\ t\ dalam\ dokumen\ d}{jumlah\ total\ kata\ dalam\ dokumen\ d} \quad (1)$$

Frekuensi jumlah kata yang muncul dalam sebuah dokumen dikenal sebagai Term Frequency (TF). Nilai TF biasanya dibagi dengan panjang dokumen (jumlah kata lengkap dalam dokumen). Ini disebabkan oleh variabel panjang dokumen dan terkait rumus IDF yaitu :

$$idf_d = \log\left(\frac{Number\ of\ document}{Number\ of\ document\ with\ term\ t'}\right) \quad (2)$$

Dimana N adalah jumlah dokumen dalam kumpulan dokumen dan n adalah jumlah dokumen yang mengandung kata tersebut. Setelah nilai TF dan IDF ditentukan, maka nilai IDF. Kata-kata dengan nilai TF-IDF tinggi dianggap penting dan lebih berkontribusi dalam menentukan topik suatu dokumen atau kumpulan dokumen.

Setelah mendapatkan TF dan IDF, kemudian dapat menghitung nilai TF-IDF yang merupakan hasil perkalian dari TF dan IDF. Rumus yang umum digunakan sebagai berikut.

$$tfidf_{t,d} = tf_{t,d} \times idf_d \quad (3)$$

Keterangan :

t	= Kata kunci term
d	= Dokumen
t.d	= Nilai TF-IDF untuk kata t dalam dokumen d
Tf	= Banyaknya t (kata) yang dicari dalam dokumen
Idf	= Banyak t kebalikan dari kata yang dicari

2.5. Oversampling

Untuk mengatasi ketidakseimbangan kelas dalam data, teknik oversampling sintetis minoritas SVM-SMOTE digunakan. Ketidakseimbangan kelas terjadi ketika setiap kelas memiliki kelas mayoritas dan minoritas. Tidak memperhatikan ketidakseimbangan ini dapat menyebabkan model memiliki bias yang sangat besar terhadap kelas mayoritas, yang dapat membuat model tidak sensitif terhadap hal-hal yang berkaitan dengan kelas minoritas yang mungkin memiliki nilai prediktif yang signifikan. Karena kelas mayoritas berkuasa, akurasi model dapat sangat tinggi, tetapi hasil prediksi untuk kelas minoritas akan sangat rendah.[11]

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

2.6. Split Data

Pembagian dataset menjadi data latih dan data uji merupakan langkah penting dalam proses pengembangan model sentimen. Data latih digunakan untuk membangun model, sementara data uji berfungsi untuk mengevaluasi kinerjanya. Penelitian ini menerapkan rasio pembagian 80:20, mengikuti studi sebelumnya yang menggunakan algoritma SVM dan pembobotan TF-IDF. Pilihan rasio ini terbukti menghasilkan akurasi tertinggi dalam penelitian terdahulu. Hasil yang diperoleh menunjukkan performa yang sangat baik, dengan rata-rata akurasi mencapai 94.00%, precision 95.32%, recall 95.05%, dan F1-Score 95.18%. Angka-angka ini mengindikasikan bahwa model yang dikembangkan memiliki kemampuan yang sangat baik dalam mengklasifikasikan sentiment [12].

2.7. Klasifikasi

Support Vector Machine (SVM), sebuah metode klasifikasi berbasis machine learning, memprediksi kelas berdasarkan model atau pola yang dihasilkan dari proses pelatihan. Penelitian ini mengadopsi SVM sebagai metode klasifikasi utama, mengingat keunggulannya yang berakar pada teori pembelajaran statistika dan performanya yang sering melampaui metode lain. Proses klasifikasi SVM melibatkan tahap pelatihan yang menghasilkan pola atau nilai acuan, dilanjutkan dengan tahap pengujian. Pada tahap pengujian, SVM menerapkan pola yang telah dipelajari untuk memberikan label sentimen pada tweet, sekaligus menghasilkan tingkat akurasi klasifikasi. Dengan demikian, SVM tidak hanya mampu mengklasifikasikan data baru, tetapi juga memberikan ukuran kehandalan hasil klasifikasinya [13]. Dibawah merupakan rumus umum Support Vector Machine (SVM):

$$f(x) = sign(w.x + b) \quad (4)$$

Keterangan :

- f(x) : fungsi prediksi yang menentukan label kelas dari input x
- sign : fungsi tanda yang mengembalikan +1 jika argumen positif dan -1 jika argumen negatif. Dalam konteks SVM, fungsi ini digunakan untuk menentukan ke kelas mana data x akan diklasifikasikan.
- w adalah vektor bobot yang menentukan arah dan orientasi hyperplane.
- x merupakan vektor fitur dari data input
- b adalah bias atau intercept dari hyperplane.

Pada metode SVM pada Penelitian ini menggunakan kernel. Kernel digunakan untuk mengelompokan data dari dimensi kecil menjadi dimensi besar. Kernel Linear adalah jenis kernel yang paling sederhana dan cepat dalam SVM, ideal untuk data yang dapat dipisahkan secara linear. Kernel ini menghitung produk dalam dari dua vektor fitur, bekerja sangat baik dengan dataset besar dan banyak fitur, namun kurang efektif untuk data non-linear [14].

2.8. Evaluasi

Tahap evaluasi bertujuan mengukur akurasi model yang telah dilatih pada dataset pelatihan. Proses ini menggunakan Confusion Matrix untuk menganalisis hasil klasifikasi algoritma Support Vector Machine (SVM). Evaluasi dilakukan dengan membandingkan hasil dari dua set data yang berbeda, serta menghitung empat metrik utama: precision, recall, F1-score, dan accuracy. Metode ini memungkinkan penilaian komprehensif terhadap performa model dalam mengklasifikasikan data, memberikan gambaran yang jelas tentang efektivitas SVM dalam tugas klasifikasi yang diberikan [15].

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

Dalam evaluasi model klasifikasi machine learning, accuracy menjadi salah satu metrik penting untuk mengukur performa model. Metrik ini menghitung persentase prediksi yang tepat dari keseluruhan prediksi yang dihasilkan oleh model. Accuracy dapat dihitung menggunakan Confusion Matrix, yang menyediakan formula khusus untuk menentukan tingkat keakuratan prediksi model klasifikasi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \; x \; 100\% \quad (5)$$

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Pengumpulan data dilakukan melalui Twitter dengan efisien, diikuti oleh analisis sentimen, tren, dan topik populer menggunakan Google Collab. Platform ini menyediakan lingkungan pemrograman berbasis cloud yang kompatibel dengan Python, memungkinkan penggunaan pustaka untuk mengakses API Twitter. Hasil pengumpulan data tweet pada file csv terdiri beberapa kolom yaitu :

1. Created_at
2. Full_text
3. Reply_count
4. Retweet_count
5. Favorite_count
6. Lang
7. User_id_str
8. Username
9. Tweet_url

Tabel 1 akan memvisualisasikan beberapa kolom yang terdapat pada file csv.

Tabel 3 Hasil Pengumpulan Data

created_at	full_text	username
Wed Mar 20 15:38:52 +0000 2024	@Naz_lira Yg perlu dijadikan pertimbangan bg siapapun yg ingin menetap di IKN adalah suku asli Kalimantan bersifat terbuka namun jk pendatang melampaui batas mk mereka tanpa kompromi akan melakukan hal layaknya Tragedi Sampit.	fadillahyuri
Wed Mar 20 15:37:37 +0000 2024	@PaimoMontok @Dragonking_03 @Siantar72 @DJoker_GinOngs @andrikosasih09 @AriHend34760216 @her_ermi Sebentar lagi IKN tanah longsor dan banjir bandang	Cintaci33333158
...	...	...
Wed Mar 20 13:31:40 +0000 2024	Investor Semakin Bertambah Aguan Bawa Geng Konglomerat Buat Investasi di IKN #jatimdamai 40 M #HiddenLove Solaria Trisakti Tangsel #PLEDIS_PROTECT_JOSHUA shopee fruity PRESAVE ROCKSTAR https://t.co/fKKw04Gv2I	Es_kopyor_real
Wed Mar 20 13:30:52 +0000 2024	Bendungan Sepaku Semoi Sudah Terisi Air dan Siap Menyuplai Kebutuhan Air Bersih di IKN #jatimdamai 40 M #HiddenLove	TeropongSakti

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

Proses ini melibatkan pengumpulan tweet berdasarkan kata kunci "Pembangunan IKN" untuk analisis lebih lanjut. Crawling data dilaksanakan selama 10 menit, data yang berhasil dikumpulkan sebanyak 1027 data tweet perihal topik pemindahan Ibu Kota Indonesia.

3.2. Labelling Data

Dalam penelitian ini, sampel data yang digunakan terdiri dari 1027 tweet yang dikumpulkan melalui proses crawling dari platform Twitter. Setiap tweet dipilih berdasarkan kriteria memiliki satu aspek spesifik terkait topik penelitian. Gambar 2 akan menampilkan grafik presentase sentimen.

3.4 Pre-Processing

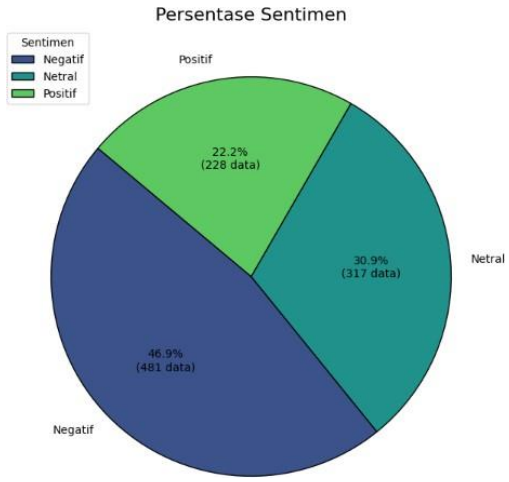
Sebelum data tweet dapat digunakan ke tahap selanjutnya, dilakukan tahap pre-processing yang bertujuan untuk membersihkan data. Tahap ini terdiri dari cleaning, tokenizing, case folding, stopword removal dan stemming. Pada gambar 3 akan menampilkan teks sebelum dan sesudah dilakukan pre-processing.

Gambar 11 Wordcloud Sebelum dan Sesudah Pre-processing

Gambar 3 menunjukkan gambaran menyeluruh tentang proses pre-processing, menunjukkan bahwa data yang telah diolah selama tahap ini masih memiliki beberapa kekurangan. Terutama pada tahap stopword removal di mana kata-kata atau kalimat yang disingkat, seperti kata-kata seperti "yg", "ga", "gak", "kalo", dan sebagainya, tidak dapat dihilangkan. Masalah tersebut akan berpengaruh pada Tahap pembobotan TF-IDF karena terdapat kata yang sering muncul namun tidak dapat dihilangkan.

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]



Gambar 12 Diagram Hasil Pelabellan

Hasil analisis sentimen yang ditampilkan menunjukkan distribusi yang tidak seimbang antar kategori. Sentimen negatif mendominasi dengan proporsi terbesar yaitu 46.9% dari total data. Ini kontras dengan sentimen positif yang hanya mencapai 22.2%, sementara sentimen netral berada di tengah dengan 30.9%. Ketidakseimbangan ini terlihat jelas dalam jumlah absolut tweet untuk setiap kategori sentimen. Data tweet bersentimen negatif berjumlah 481 jauh melebihi jumlah tweet bersentimen positif yang hanya 228. Perbedaan signifikan ini mengindikasikan adanya ketimpangan dalam distribusi data sentimen, dengan kecenderungan kuat ke arah sentimen negatif dalam dataset yang dianalisis.

3.4 Pre-Processing

Sebelum data tweet dapat digunakan ke tahap selanjutnya, dilakukan tahap pre-processing yang bertujuan untuk membersihkan data. Tahap ini terdiri dari cleaning, tokenizing, case folding, stopword removal dan stemming. Pada gambar 3 akan menampilkan teks sebelum dan sesudah dilakukan pre-processing.

Gambar 13 Wordcloud Sebelum dan Sesudah Pre-processing

Gambar 3 menunjukkan gambaran menyeluruh tentang proses pre-processing, menunjukkan bahwa data yang telah diolah selama tahap ini masih memiliki beberapa kekurangan. Terutama pada tahap stopword removal di mana kata-kata atau kalimat yang disingkat, seperti kata-kata seperti "yg", "ga", "gak", "kalo", dan sebagainya, tidak dapat dihilangkan. Masalah tersebut akan berpengaruh pada Tahap pembobotan TF-IDF karena terdapat kata yang sering muncul namun tidak dapat dihilangkan.

3.5 Pembobotan Kata TF-IDF

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]

Setelah mendapatkan data bersih atau "Tweet_clean" yang disimpan pada file HasilPreproFinal.csv, langkah selanjutnya adalah mengubah kata menggunakan fungsi TF-IDF. Pada gambar 4 dari dokumen pertama (indeks 0) Dimana term 2475 memiliki hasil TF-IDF sekitar 0.2519, perhitungan TF-IDF mengubah kata pada dokumen terakhir (indeks 1026), di mana term yang diindeks pada 1070 dalam kosakata memiliki skor TF-IDF sekitar 0.0528, kosakata adalah kamus yang memiliki indeks untuk setiap kata.

```
(0, 2475)     0.25193917032228064
(0, 2928)     0.25193917032228064
(0, 1566)     0.23783403448386145
(0, 1534)     0.18424573275412948
(0, 1446)     0.25193917032228064
(0, 1802)     0.21372114550203783
(0, 331)      0.22006366492830803
(0, 1541)     0.25193917032228064
(0, 625)      0.1627108782148137
(0, 1216)     0.227826281340457
(0, 496)      0.2083586217359531
(0, 2620)     0.23783403448386145
(0, 1277)     0.1513902316999714
(0, 226)      0.19263516874369266
(0, 2737)     0.21372114550203783
(0, 1070)     0.03627397559204956
(0, 2879)     0.17960050337680972
(0, 2616)     0.20371339235863334
(0, 402)      0.25193917032228064
(0, 2895)     0.23783403448386145
(0, 1151)     0.11584949258886452
(0, 2192)     0.17362225476960494
(0, 3120)     0.19685749909530723
(1, 292)      0.4727705155420468
(1, 307)      0.39680007953411295
:     :
(1025, 451)   0.22492135459801635
(1025, 949)   0.20063839100313716
(1025, 665)   0.20491792800540193
(1025, 2826)  0.20723959039943615
(1025, 805)   0.2776956749616
(1025, 1209)  0.21512582598612218
(1025, 201)   0.21232347118957007
(1025, 1813)  0.20491792800540193
(1025, 2329)  0.2027218237761171
(1025, 2513)  0.21512582598612218
(1025, 1070)  0.084707187997168895
(1026, 3048)  0.3667703720204228
(1026, 2116)  0.3033262716154931
(1026, 61)    0.3111329768377866
(1026, 698)   0.2965637959565459
(1026, 1393)  0.2804362355165412
(1026, 1899)  0.2804362355165412
(1026, 2184)  0.25839027461610514
(1026, 2975)  0.2349614745044527
(1026, 2177)  0.25275743700915504
(1026, 706)   0.29059886968708887
(1026, 1091)  0.1980164075196778
(1026, 1725)  0.1825508609452782
(1026, 306)   0.29291553859560593
(1026, 1070)  0.05280726893534258
```

Gambar 14 Hasil TF-IDF

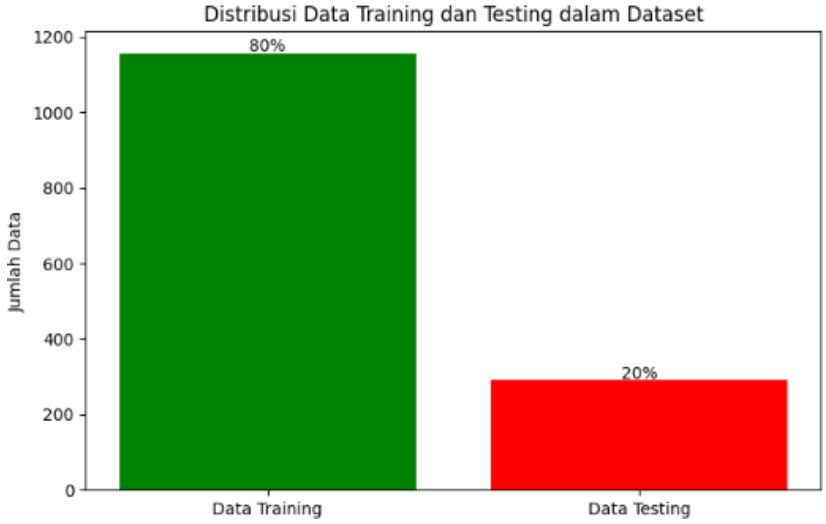
Hasil pembobotan kata yang dilakukan dengan menggunakan ekstraksi fitur TF-IDF menunjukkan bahwa kata kunci tertentu memiliki pengaruh yang signifikan pada pertimbangan sentimen yang berkaitan dengan pemindahan Ibu Kota Indonesia. "IKN", "bangun", "Kalimantan", "pindah", dan "Investor" adalah kata-kata yang paling sering dibicarakan dalam tweet pengguna.

3.6 Hasil Split Data

Setelah mengetahui hasil pembobotan TF-IDF secara otomatis, langkah berikutnya adalah membagi secara acak pada file.csv yang akan diolah dengan perbandingan 80:20. Di mana 20% data pengujian dan 80% data pelatihan. Seringkali, metode ini dipilih untuk menyediakan jumlah data yang cukup untuk model untuk belajar (set pelatihan) dan juga menyisihkan jumlah data yang cukup untuk mengevaluasi kinerja model. Split data akan ditampilkan pada gambar 5.

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

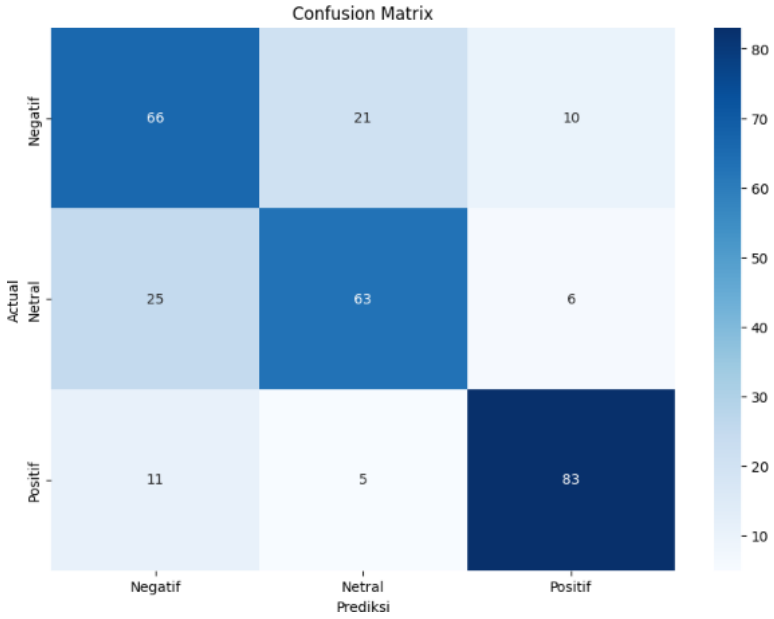
JURNAL TEKNOLOGI INFORMASI
Jurnal Keilmuan dan Aplikasi Bidang Teknik Informatika
[E-ISSN 2656-0321]
[Vol 18 No 2]
[Agustus 2024]



Gambar 15 Hasil Split Data

3.7 Klasifikasi dan Evaluasi

Keputusan untuk menggunakan kernel linear dalam eksperimen didasarkan pada konteks dan kebutuhan praktis, bukan hanya hasil numerik. Untuk Support Vector Machine, kernel linear memiliki banyak keunggulan, terutama dalam klasifikasi data yang berukuran besar dan data dapat dipisahkan secara linear dan mengurangi risiko overfitting dibandingkan dengan kernel yang lebih kompleks seperti Polinomial atau RBF, terutama untuk dataset yang tidak terlalu besar [16]. Untuk Tampilan Hasil *Confusion Matrix* akan ditampilkan pada gambar 6.



Gambar 16 Hasil *Confusion Matrix*

Pada hasil Confusion Matrix prediksi dan actual, model berhasil memprediksi data sebenarnya. Pada baris actual Negatif dan prediksi Negatif memiliki 66 data, actual Netral prediksi Netral memiliki 63 data dan actual Positif prediksi Positif memiliki 83 data. Sisanya terdapat model yang salah meprediksi data

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI

yaitu 21 data actual negatif yang diprediksi netral, 10 data actual negatif diprediksi positif, 25 data actual netral yang diprediksi negatif, 6 data actual netral diprediksi positif dan 11 data actual positif diprediksi negative, 5 data actual positif diprediksi netral.

Akurasi mode SVM mencapai 73% hasil tersebut menunjukan bahwasannya metode Support Vector Machine (SVM) dan pembobotan kata menggunakan ekstraksi fitur TF-IDF dalam klasifikasi sentiment perihal pemindahan Ibu Kota Indonesia menghasilkan akurasi yang cukup baik.

4. Kesimpulan

Penelitian yang telah dilakukan menunjukkan bahwa metode klasifikasi yang menggunakan algoritma Support Vector Machine (SVM) dan ekstraksi fitur TF-IDF pada data sentiment pembangunan IKN efektif digunakan pada penelitian ini. Hasil evaluasi model menggunakan Matriks Konflik menunjukkan bahwa akurasi klasifikasi SVM pada data tweet pembangunan IKN mencapai 73%. Ini menunjukkan bahwa metode SVM dengan pembobotan TF-IDF mampu mengklasifikasikan data sentiment dengan hasil yang cukup baik.

Namun memiliki beberapa kekurangan seperti kata-kata yang belum dihapus pada tahap stoword contohnya seperti kata "yg", "ga", "gak", "jd", dll., yang dapat mempengaruhi pembobotan kata karena kata-kata ini sering muncul di dokumen. Untuk penelitian selanjutnya disarankan pada tahap stopword removal sebaiknya mengelola dan menghapus secara manual maupun secara otomatis menggunakan library sastrawi. Untuk penelitian selanjutnya, eksplorasi metode dan algoritma selain SVM. Terdapat beberapa penelitian sebelumnya yang sama membahas topik IKN dengan menggunakan model klasifikasi berbeda, seperti pada penelitian [17] menggunakan model klasifikasi Naïve Bayes, terdapat juga penelitian [18] dengan model klasifikasi menggunakan K-NN dan klasifikasi model menggunakan metode Random Forest dan Logistic Regression seperti penelitian berikut [19]

Daftar Pustaka

[1] W. Liano Hutasoit, “ANALISA PEMINDAHAN IBUKOTA NEGARA”.

[2] M. K. Saraswati *et al.*, “Pemindahan Ibu Kota Negara Ke Provinsi Kalimantan Timur Berdasarkan Analisis Swot,” *Jurnal Ilmu Sosial dan Pendidikan (JISIP)*, vol. 6, no. 2, pp. 2598–9944, 2022, doi: 10.36312/jisip.v6i1.3086/http.

[3] S. D. Saputra, T. Gabriel J, and M. Halkis, “ANALISIS STRATEGI PEMINDAHAN IBU KOTA NEGARA INDONESIA DITINJAU DARI PERSPEKTIF EKONOMI PERTAHANAN (STUDI KASUS UPAYA PEMINDAHAN IBU KOTA NEGARA DARI DKI JAKARTA KE KUTAI KARTANEGARA DAN PENAJAM PASER UTARA) STRATEGY ANALYSIS RELOCATION OF THE CAPITAL CITY OF INDONESIA FROM DEFENSE ECONOMIC PERSPECTIVE (CASE STUDY OF RELOCATION OF THE CAPITAL CITY FROM DKI JAKARTA TO KUTAI KARTANEGARA AND PENAJAM PASER UTARA),” 2021.

[4] D. Oktavia and Y. R. Ramadahan, “Analisis Sentimen Terhadap Penerapan Sistem E-Tilang Pada Media Sosial Twitter Menggunakan Algoritma Support Vector Machine (SVM),” *Media Online)*, vol. 4, no. 1, pp. 407–417, 2023, doi: 10.30865/klik.v4i1.1040.

[5] N. Legiawati, T. I. Hermanto, and Y. R. Ramadhan, “Analisis Sentimen Opini Pengguna Twitter Terhadap Perusahaan Jasa Ekspedisi Menggunakan Algoritma Naïve Bayes Berbasis PSO,”

JURIKOM (Jurnal Riset Komputer), vol. 9, no. 4, p. 930, Aug. 2022, doi: 10.30865/jurikom.v9i4.4629.

[6] V. Fitriyana *et al.*, "Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine," 2023.

[7] H. C. Husada and A. S. Paramita, "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM)," *Teknika*, vol. 10, no. 1, pp. 18–26, Feb. 2021, doi: 10.34148/teknika.v10i1.311.

[8] O. Zoellanda ATane, K. Muslim Lhaksmana, and F. Nhita, "Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM (Support Vector Machine)."

[9] O. I. Gifari, M. Adha, I. Rifky Hendrawan, F. Freddy, and S. Durrand, "Analisis Sentimen Review Film Menggunakan TF-IDF dan Support Vector Machine," *JIFOTECH (JOURNAL OF INFORMATION TECHNOLOGY*, vol. 2, no. 1, 2022.

[10] J. Muliawan and E. Dazki, "SENTIMENT ANALYSIS OF INDONESIA'S CAPITAL CITY RELOCATION USING THREE ALGORITHMS: NAÏVE BAYES, KNN, AND RANDOM FOREST," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 5, pp. 1227–1236, 2023, doi: 10.52436/1.jutif.2023.4.5.347.

[11] M. Persada Pulungan, A. Purnomo, A. Kurniasih, S. Tinggi Ilmu Manajemen dan Ilmu Komputer ESQ, and P. Korespondensi, "PENERAPAN SMOTE UNTUK MENGATASI IMBALANCE CLASS DALAM KLASIFIKASI KEPRIBADIAN MBTI MENGGUNAKAN NAIVE BAYES CLASSIFIER APPLICATION OF SMOTE TO OVERCOME CLASS IMBALANCE IN THE MBTI PERSONALITY CLASSIFICATION USING THE NAÏVE BAYES CLASSIFIER," vol. 10, no. 7, pp. 1493–1502, 2023, doi: 10.25126/jtiik.2023107989.

[12] M. F. Asshiddiqi and K. M. Lhaksmana, "Perbandingan Metode Decision Tree dan Support Vector Machine untuk Analisis Sentimen pada Instagram Mengenai Kinerja PSSI."

[13] A. Novantirani, M. S. Kania Sabariah, and V. Effendy, "Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine."

[14] S. Rahayu and Y. Yamasari, "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)," *Journal of Informatics and Computer Science*, vol. 05, 2024.

[15] R. Tineges, A. Triayudi, and I. D. Sholihati, "Analisis Sentimen Terhadap Layanan Indihome Berdasarkan Twitter Dengan Metode Klasifikasi Support Vector Machine (SVM)," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 3, p. 650, Jul. 2020, doi: 10.30865/mib.v4i3.2181.

[16] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review," *Information (Switzerland)*, vol. 15, no. 4, Apr. 2024, doi: 10.3390/info15040235.

[17] K. A. Lubis, M. Theo, A. Bangsa, and A. Yudertha, "ANALISIS SENTIMEN OPINI MASYARAKAT TERHADAP PINDAHNYA IBU KOTA INDONESIA DENGAN MENGGUNAKAN KLASIFIKASI NAÏVE BAYES," 2024. [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

[18] O. Muhammad and I. Ramadhon, "ANALISIS SENTIMEN TERHADAP PEMINDAHAN IBU KOTA INDONESIA PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE ALGORITMA K-NEAREST NEIGHBOR (K-NN) SKRIPSI."

[19] M. Putri Agustina, "Sentimen Masyarakat Terkait Perpindahan Ibukota Via Model Random Forest dan Logistic Regression," *AITI: Jurnal Teknologi Informasi*, vol. 18, no. Agustus, pp. 111–124, 2021.

https://doi.org/10.47111/JTI
Available online at https://e-journal.upr.ac.id/index.php/JTI